REPORT DOCUMENTATION PAGE

2

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
27-08-2014	Final Report	1-Feb-2011 - 31-Jan-2014

4. TITLE AND SUBTITLE

Final Report: Hybrid Josephson Junction MRAM Cryogenic Memory

5a. CONTRACT NUMBER

W911NF-11-1-0073

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHORS

Gerald W. Gibson, Jr.

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES

IBM Corporation (T.J. Watson Research Lal
1101 Kitchawan Road

Yorktown Heights, NY 10598 -0000

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES)

U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
59462-PH.1

12. DISTRIBUTION AVAILIBILITY STATEMENT

Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT

This report identifies a candidate architecture for an energy efficient SFQ-based processor unit. It quantitatively analyzes candidates for register and main memory with respect to access time, density and energy efficiency. Finally, it investigates the efficacy of three dimensional integration (3DI) for enhancing aerial circuit density.

15. SUBJECT TERMS

RSFQ, Cryogenic Memory, Energy Efficient, 3DI

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Gerald Gibson
UU	UU	UU	UU		19b. TELEPHONE NUMBER: 914-945-1206

Report Title

Final Report: Hybrid Josephson Junction MRAM Cryogenic Memory

ABSTRACT

This report identifies a candidate architecture for an energy efficient SFQ-based processor unit. It quantitatively analyzes candidates for register and main memory with respect to access time, density and energy efficiency. Finally, it investigates the efficacy of three dimensional integration (3DI) for enhancing aerial circuit density.

Enter List of papers submitted or published that acknowledge ARO support from the start of the project to the date of this printing. List the papers, including journal references, in the following categories:

(a) Papers published in peer-reviewed journals (N/A for none)

Received Paper

TOTAL:

Number of Papers published in peer-reviewed journals:

(b) Papers published in non-peer-reviewed journals (N/A for none)

Received Paper

TOTAL:

Number of Papers published in non peer-reviewed journals:

(c) Presentations

Number of Presentations: 0.00

Non Peer-Reviewed Conference Proceeding publications (other than abstracts):

Received Paper

TOTAL:

Number of Non Peer-Reviewed Conference Proceeding publications (other than abstracts):

Peer-Reviewed Conference Proceeding publications (other than abstracts):

Received Paper

TOTAL:

Number of Peer-Reviewed Conference Proceeding publications (other than abstracts):

(d) Manuscripts

Received Paper

TOTAL:

Number of Manuscripts:

Books

Received Book

TOTAL:

Received Book Chapter

TOTAL:

Patents Submitted

Patents Awarded

Awards

Graduate Students

NAME	PERCENT_SUPPORTED
FTE Equivalent:	
Total Number:	

Names of Post Doctorates

NAME	PERCENT_SUPPORTED
FTE Equivalent:	
Total Number:	

Names of Faculty Supported

NAME	PERCENT_SUPPORTED
FTE Equivalent:	
Total Number:	

Names of Under Graduate students supported

NAME	PERCENT_SUPPORTED
FTE Equivalent:	
Total Number:	

Student Metrics

This section only applies to graduating undergraduates supported by this agreement in this reporting period

The number of undergraduates funded by this agreement who graduated during this period: 0.00

The number of undergraduates funded by this agreement who graduated during this period with a degree in science, mathematics, engineering, or technology fields:..... 0.00

The number of undergraduates funded by your agreement who graduated during this period and will continue to pursue a graduate or Ph.D. degree in science, mathematics, engineering, or technology fields:..... 0.00

Number of graduating undergraduates who achieved a 3.5 GPA to 4.0 (4.0 max scale):..... 0.00

Number of graduating undergraduates funded by a DoD funded Center of Excellence grant for Education, Research and Engineering:..... 0.00

The number of undergraduates funded by your agreement who graduated during this period and intend to work for the Department of Defense 0.00

The number of undergraduates funded by your agreement who graduated during this period and will receive scholarships or fellowships for further studies in science, mathematics, engineering or technology fields: 0.00

Names of Personnel receiving masters degrees

NAME

Total Number:

Names of personnel receiving PHDs

NAME

Total Number:

Names of other research staff

NAME PERCENT_SUPPORTED

FTE Equivalent:
Total Number:

Sub Contractors (DD882)

Inventions (DD882)

Scientific Progress

Technology Transfer

Final Report: Computer Architecture for Energy Efficient SFQ

ARO Grant W911NF-11-1-0073
PI: Gerald Gibson
August 27, 2014

This report summarizes the work accomplished during this ARO-sponsored project at IBM Research to identify and model an energy efficient SFQ-based computer architecture.

The promise of energy efficient SFQ logic is summarized in Figure 1. The plot compares the energy per logical operation of 11nm CMOS to that of zero quiescent power variants of SFQ as a function of clock frequency and SFQ system scale. The reason for the multiple SFQ curves derives from the self-consistent treatment of refrigeration power overhead and its improved efficiency with heat extraction capacity. In the plot,

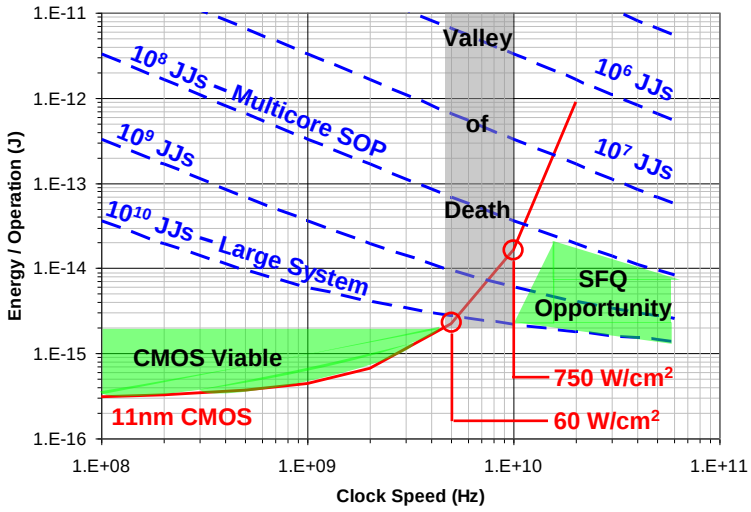

Figure 1: Energy per Logical Operation vs. Clock Frequency

a factor of ten improvement in energy efficiency, relative to CMOS, has been arbitrarily chosen as the criterion for viability of SFQ. The analysis suggests that the scale of the SFQ-based system will have to be 100 million junctions, minimum, which is 3-4 orders of magnitude greater than any superconducting electronics circuit which has been successfully built to date. On a more positive note, it appears that a relatively modest clock speed of 10-20 GHz is sufficient for differentiation.

The first six months of this program were spent carefully evaluating architectures and underlying technologies and how they fit together. This work comprised harvesting the wealth of architectural infrastructure that exists for silicon-based (or more generically voltage-state) logic while remaining cognizant of the fact that these architectures have successfully evolved *because* they exploit the basic nature of the devices of which they are comprised. Because SFQ logic, which encodes data as the presence or absence of a quantum of magnetic flux, differs in very fundamental ways from voltage state logic, it

may well be that it is ill suited for implementation in "off the shelf" architectures. Ultimately, an approach was chosen which allows incremental development of SFQ-based architecture.

The constraints imposed by requirements of the memory subsystem dominate all other considerations. It is readily apparent that the sophisticated memory hierarchies of x86 architectures, even in the case of older single-core technologies such as the Pentium III, place requirements on both latency and density of memory bits which can not be met by any existing cryogenic memory technology, nor by any which might be realized within an intermediate time horizon. This observation led to consideration of simpler computing engines which might be scaled out into large parallel systems in order to meet device count requirements dictated by cooler efficiency. One such architecture is IBM Windsor Blue (WB), illustrated schematically in Figure 2. The basic building block of WB is a "tile" comprised of a 64-bit arithmetic logic unit with 256 64-bit registers adjacent to the ALU. The transistor count for the ALU and memory control is modest, only about 500,000. Assuming rough parity between transistor and Josephson junction counts, this constitutes a chip which is complex by historical superconducting logic standards, but not untenable. It is assumed that multilevel metallization will allow the roughly 16k bits of registers to be effectively "stacked" over the ALU and memory control circuitry. Because of the poor device density of SFQ, relative to CMOS, the

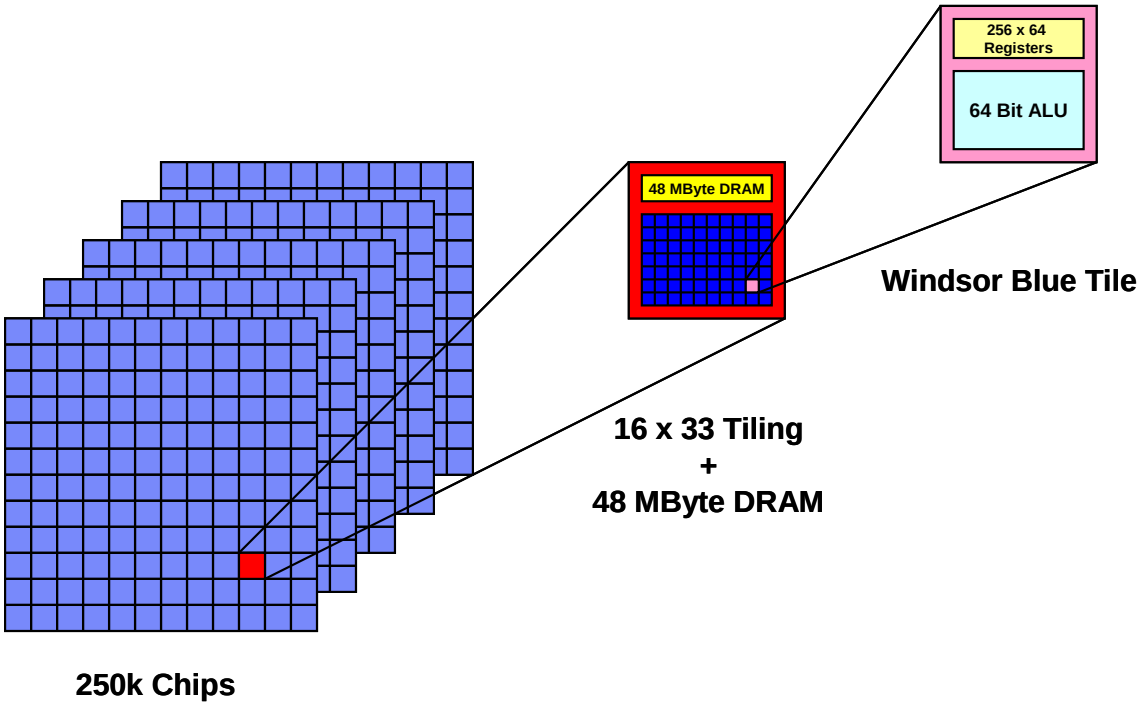

Figure 2: The IBM Windsor Blue System

number of WB tiles on a chip will only be of order 10; however, a main memory of cryogenic CMOS or MRAM, potentially stacked in a plane above or below the Josephson

devices, would meet the RAM per tile requirement. Table 1 provides an approximate sizing of the SFQ-based WB chip based on optimistic, but not unreasonable, assumptions regarding the size and performance of devices, all of which have been demonstrated, at least at rudimentary levels

The term "clock speed", used above, requires clarification. In CMOS, the clock sets and releases latches. In between these latches are layers of sequential logic, often as much as 20 gates deep. This stands in stark contrast to traditional SFQ, in which each logic gate is clocked. In a comparison of identically architected circuits, this means that the SFQ circuit must be clocked at a rate that is higher than the CMOS circuit by roughly a factor of the mean number of logic levels between latches in order to reach parity processing speed. To circumvent this, architectures with extreme pipelining, and concomitant complexity and timing challenges, have been proposed. Unfortunately, both leading energy efficient variants of SFQ logic require clocked gates, and will therefore either need to operate at clock speeds in the 50-100 GHz range, or adopt fairly radical processor architectures in order to compete with mainstream CMOS.

Tile

- Assume 11 square microns per JJ including passives overhead
- JJ Count: 575k (500k for ALU and memory control + 15% JTL overhead)

→ **ALU is 2.5mm / side**

Registers

- We require 16k Bits → allows 390 um^2 (20um x 20um) per bit
- VT Mem
- 16 k Bit Array --> 3.8mm x 3.7mm
  (Nagsawa et al, IEEE Trans App. Supercond. 17(2) 177)
- # JJs = 81k --> 704k um^2

→**Increases Tile size to 2.7mm / side**

JJ MRAM Main Memory

- 48M Bytes for a 16x33 array of tiles → 727k bit / tile → 10 um^2 / bit
- **Write time is ~10ns → latency of ~100 @ 10GHz**

Chip

- Room to do a 3x3 tiling on a 1cm chip
- Each chip has (3x3x10) / (16x32) ~ 1/6 computing power of CMOS WB chip

→ **Require ~ 6x SFQ chips for SC WB as CMOS WB**

Table 1: Approximate Sizing of SFQ WB Chip

Analytical Modeling of Memory Architecture

Three main classes of memory have been successfully demonstrated to operate at 4K:

- Vortex-transitional Cells
- Hybrid Josephson – CMOS
- Hybrid Josephson – MRAM

In order to proceed with modeling, it has been necessary to determine, for the Vortex Transitional and Hybrid JJ-MRAM memories, both of which are cross-point arrays, the largest memory block which can operate at a given clock speed, parameterized by memory density and size (in bits). It is common in the case of Si-based computing systems to optimize random access memories for, e.g. energy efficiency, by breaking up the largest possible block into smaller sub-blocks. The analysis which follows does not include, but neither precludes, such optimization exercises.

Vortex-transitional (VT) cells, summarized in the design study of Nagasawa *et al*[4], considered the speed and power performance of deeply pipelined random access memory arrays based on VT cells. The designs, which have never been realized, rely on eleven layers of Nb metallization in order to achieve their packing density. The performance results are shown graphically in Figure 3. The design was constrained by a maximum cross-point writable block size of 1kBit. Under this constraint, the energy efficiency of pipelining diminishes with memory size and comes at write time cost. This is similar to the case of CMOS where decode and sense operations dominate the power dissipation of the memory cells, themselves. It is also important to note that bias voltage for the cell design was set to the lowest value for which proper operation of the cell obtained: 0.1mV. At such a low voltage, the replacement of the decode circuitry with an energy efficient SFQ logic variant would only improve power consumption by a few multiples.

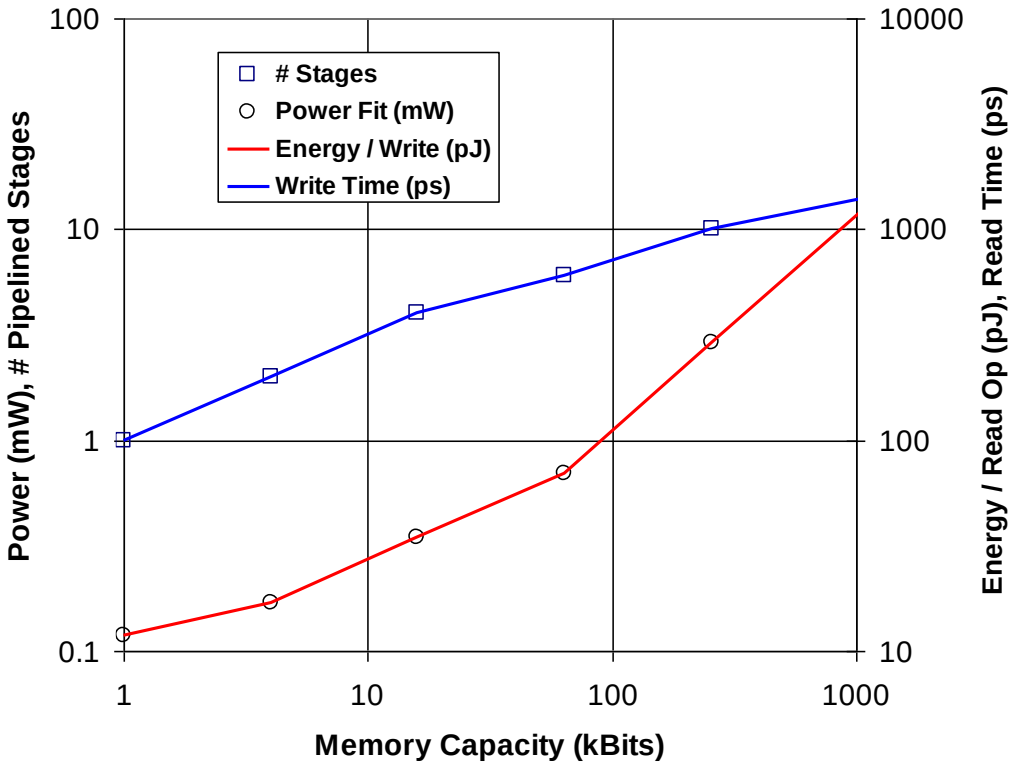

Figure 3: All Josephson Pipelined RAM

Nagasawa *et al* Supercond. Sci. Tech. v19pS325 (2006)

The deep pipelining significantly impacts the access time of the memory as a whole, since each stage requires a clock cycle to traverse. The reason given for the 1kBit maximum block size was that modeling had shown that rows of more than 32 blocks could not be driven at frequencies above the 10GHz clock frequency which had been set as a design parameter. Indeed, the cross-point writable block size constraint encountered by Nagasawa et al, is not a peculiarity of their particular driver design, but rather results from more fundamental constraints encountered when Josephson junction-based drivers are used to write flux quanta into superconducting loops used in cross-point addressable memory arrays.

Figure 4 shows a generic four-cell piece of a cross-point addressable array of single flux quantum storage loops. This calculation assumes (1) the bit is defined by the presence or absence of a single flux quantum in a superconducting loop which is written there by way of magnetic coupling to a pair of orthogonal write lines, (2) the write wires are bare, (3) for simplicity, mutual inductance between the bare write lines and the loops is ignored which means that this calculation provides lower-limits on calculated write times, (4) write time of an individual cell is of order ps and therefore much less than the L/R time constant of the bare write lines, and (5) pipelining is not considered. Referring to labeling in Figure 4, the flux storage loop, which comprises the bit, has inner dimension **a**; the separation from field wire centerline to inner loop edge, which figures into the logarithmic scaling of the coupled flux, is δ and we take the memory cell pitch to be 2a for convenience. The square loops are a natural outcome of symmetry in the cross-point write scheme. Note that the space between the inner dimensioned loop and the field lines is shown schematically to be less than that for the other three loops at the intersection, which unambiguously identifies the cell to be written and which structure can be tiled out across the block.

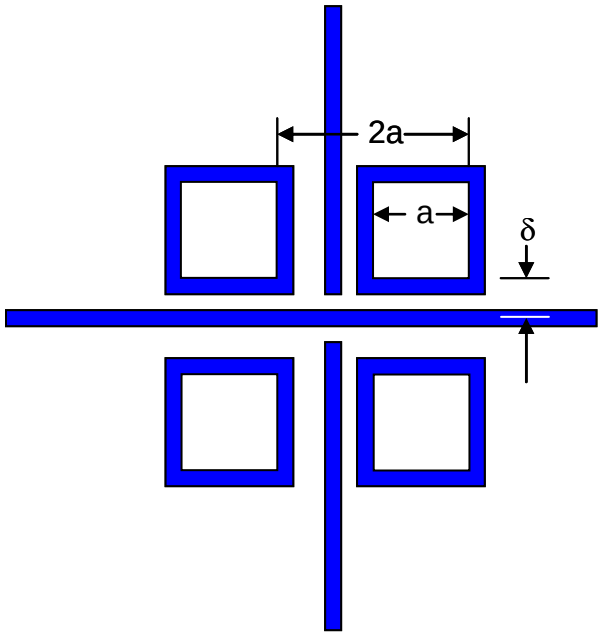

Figure 4: Schematic of Generic Flux Quantum Memory

For the purpose of quantitative calculation, the following fixed modeling parameters are assumed: write lines are 400nm thick and have critical dimension (aerial width) of 400nm, δ = 400nm, and a is allowed to vary and, thus, sets cell size.

Using this simple model with the above fixed parameters, and re-parameterizing a into the density of bits in number/cm^2, the voltage required to drive the currents of sufficient magnitude to write a flux quantum into the loop at their intersection is then calculated. The results are shown in Figure 5. The horizontal purple line represents the (arbitrary) peak drive voltage available to Josephson technology of 2mV. The plot indicates a fairly weak impact of bit density on the write line voltage: two orders of magnitude bit density change results in only a factor of two variation in drive voltage. This is because increasing the density (decreasing the loop diameter) increases the current required to write the bit, while the length of the write line (holding the array size, in bits, constant) decreases, leading to reduced inductive loading. This tradeoff between write current and line inductance represents a fundamental constraint on flux quantum storage-based memories and it is therefore no accident that this crude model yields results that are within a few factors, on the low side, of the results for the geometrically optimized cell of Nagasawa *et al*. A second critical observation is that the number of bits in a cross-point addressable block drops in a stronger-than-linear manner as a function of increasing clock speed, due to the higher voltage required to drive the (transient) current to the required level in a decreasing period of time. Finally, note that the energy for charging the bare wires increases with bit density due to its quadratic dependence on current.

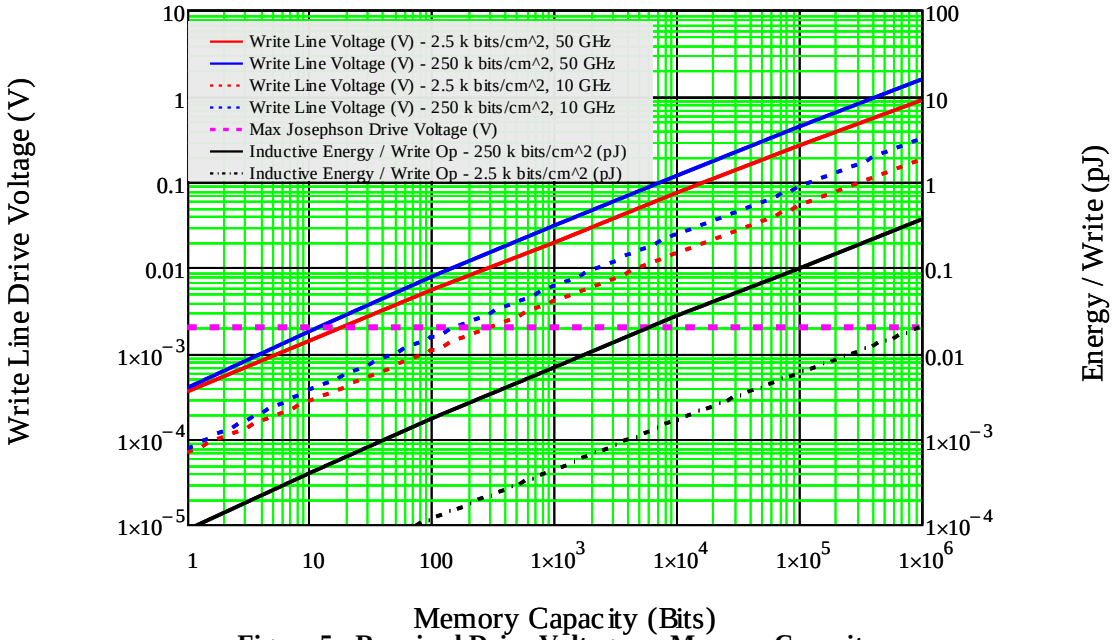

Figure 5: Required Drive Voltage vs. Memory Capacity

This long digression into the vortex transitional cell has been necessary in order to obtain meaningful comparison to the other two cryogenic memory options, MRAM and CMOS.

In our comparisons amongst the three memory candidates, we make the following assumptions:

Vortex Transitional Memory**:** Use the (optimistic) values from Table I of Nagasawa et al and assumed that each pipelined stage required a clock period.

Table 1. Performance estimation of the RAM.

RAM capacity	64-kbit	256-kbit	1-Mbit
Memory cell size (μm)	15 × 15	15 × 15	15 × 15
RAM size (mm)	4.4 × 4.4	8.9 × 8.9	18 × 18
Clock frequency (GHz)	10	10	10
Number of pipeline stages	6	10	14
Voltage of dc-power bus (mV)	0.1	0.1	0.1
Power dissipation (mW)	0.7	3	12

Figure 6: Table I from Nagasawa et al[4]

Hybrid JJ CMOS:

- Base data is taken from IEEE Transactions on Applied Superconductivity "64-kb Hybrid Josephson-CMOS 4 Kelvin RAM with 400 ps Access Time and 12 mW Read Power" Van Duzer *et al*, IEEE Trans. App. Superconductivity v23n3p1700504 (2013)
- Scaling to the 22nm CMOS node was calculated
    - Power reduction in decode logic and driver electronics by area ratio to 65nm technology
    - No improvement in decode logic and drive electronics speed
- Read and write times followed the logarithmic scaling with memory size generally applied in analytical memory modeling[6]

Hybrid Josephson-MRAM:

- Best-case cell parameters of 2μm^2 area, 1 mA write currents and 2ns MTJ write time
- These parameters yield a drive write voltage of 2mV for a 1 MBit array
- The 2mV drive voltage limit applied to the particular read scheme employed in the IBM cell allowed a maximum array size of 20kBit under the constraint of read time / write time parity. (This does not pose a serious issue, however, since pipelining can be accomplished at the timescale of the SFQ decode and drive electronics which is orders of magnitude faster than 2 ns)
- **Decode and drive contributions to power are not included, as optimal MRAM memory organization is currently under investigation → *expect 10-100x increase in power consumption***

Figures 7 and 8 compare these three cryogenic memory candidates. It is interesting to note that for memories sized for L1 cache or greater, hybrid Josephson-22nm CMOS memory competes quite favorably with all-Josephson VT-type memory in terms of both energy and speed, and is especially palatable when one considers its immense density advantage. Only when VT random access arrays drop below the 1kBit threshold does one recover the innate speed of the underlying technology. It must also be noted that

register files, which do not require random access, will be most fast and energy efficient, when comprised of all Josephson devices.

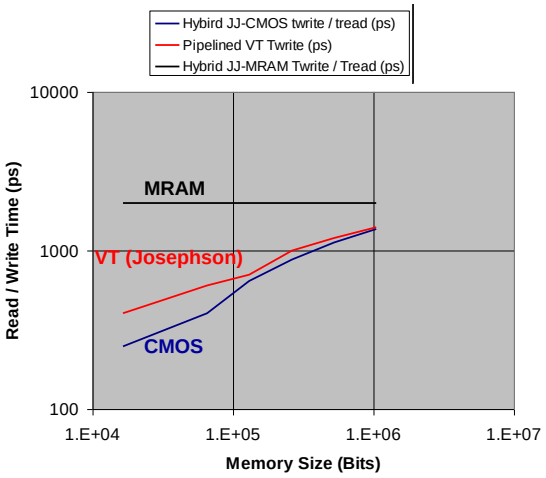

Figure 7: Access Time Comparison

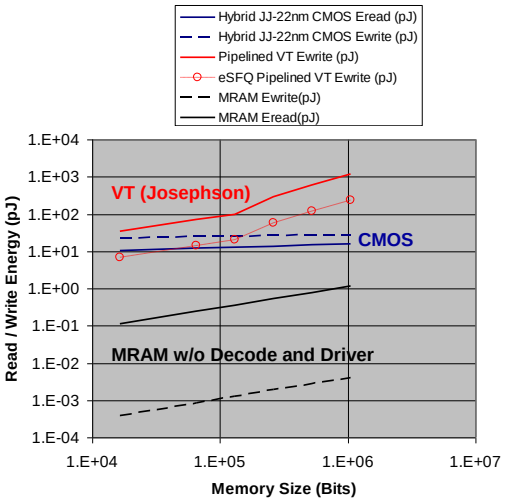

Figure 8: Energy Comparison

One assumption in the analysis of VT cell-based memory will now be revisited. It was assumed that each level of pipelining came at the cost of one clock cycle, following the exemplar NOR-gate decode of [4] and [5]. In fact, numerous schemes have been exposited for eliminating the requirement that each gate be triggered by a global clock which, thus, enables sequential logic to be employed. Released from the requirement of clocked gates, asynchronous decoders can be designed which allow pipelining that is deeper than 1 level per clock cycle. As an example, the Data Drive Self Timed (DDST) approach[7] in its simplest form requires a standard RSFQ logic gate to have appended to its output a complementary D flip flop, which costs an extra cycle of the DDST internal clock. In [7] the internal clock speed was shown to approach 40GHz in 1kA/cm^2 technology, so that in the 10kA/cm^2 technology considered by Nasagawa[4], internal clock speeds of as much as 100GHz may be realized. By it's dual rail nature, negating a DDST gate (e.g. OR → NOR) is simply a matter of swapping output rails meaning that the typical NAND- or NOR-based decoder elements would not require an additional inverter on the gate's output, so a delay reduction of as much as 5x could be possible in a system with global clock frequency of 10GHz.

The die size projections in Figure 4[4] are rather optimistic. If we consider what has actually been fabricated in an advanced six-level metal, 10kA/cm^2 technology, the picture is somewhat less encouraging. In [5] a 4kBit RAM was 2.4mm on a side, about a quarter as dense as the RAMs in [4], and it yielded at only 96.7%. Scaling the 4kBit RAM in [5] to 16kBit and assuming PTL signal propagation velocity of 10^{10} cm/s, the "Manhattan" worst case path can just be traversed in a 10 GHz clock cycle. Thus, the inherently poor density of stored flux quantum memories brings about "speed of light" constraints at relatively low bit count memories.

Advanced Packaging Techniques

Over the last decade, three dimensional integration (3DI) techniques have begun to gain traction in CMOS technology. In particular, through-silicon vias (TSV) have enabled die stacking and advanced programs have begun to look at stacking as many as eight die. TSV resistance is of order 10mΩ, which is sufficiently low that an SFQ pulse can propagate across it. Since each stacked die will be within ~1mm of, for example, a controller chip at the base of the stack, it may be possible to significantly increase the number of bits reachable in one clock cycle. Additionally, by enabling selection of known good die, the yield of large memories could be improved, as well.

The challenges associated with cryogenic 3DI are, of course, formidable. Extreme mechanical stresses resulting from thermal coefficient of expansion mis-matches will have to be managed, and careful thermal modeling and use of thermal management techniques, some potentially novel, will be required. One should expect that materials which have performed acceptably in chips or MCMs will fail under the more rigorous conditions of 3DI, and new ones will have to be developed to replace them. Architectural questions regarding memory organization will have to be dealt with, as well.
All of these considerations will be complicated by the fact that stacked chips cannot be galvanically coupled because no process exists for superconducting TSVs, as Nb cannot be electroplated in a manner compatible with Josephson technology. This is because the finite resistance of TSVs will cause resistive voltage drops along current source and return paths. The 16kBit RAM fabricated in [4] drew ~800mA, which would lead to an 8mV drop if the current were passed through a single TSV. By distributing the current across many TSVs, the level shift problem could be mitigated and, importantly, resistive power dissipation reduced. In addition, current flow within a chip, such as ground return paths, could be engineered through appropriate TSV placement. However, any shift in ground reference is unacceptable in superconducting technology. Luckily, techniques for managing signal coupling across domains with differing ground reference have been developed and demonstrated for current recycling[5,6]. In fact, because each stacked RAM chip will be identical, 3DI is a natural target for current recycling or, equivalently, serial biasing, since the current requirements for each stacked chip will be inherently balanced.

In order to interpret the results of calculations of TSV performance in a concrete and meaningful way, it is necessary to more specifically define the architecture of the processor under consideration. For the purposes of this report, we will assume a bit-serial architecture in which each processor bit addresses a register bit-slice which is 128 bits deep. The register structure is taken to be that shown in Figure 9. Main memory is comprised of the CMOS variety described in Section 2, above, and would reside on a chip to which the SFQ components are attached, either as part of an MCM, or vertically stacked with TSV connections. The approximate junction count of the processor bit-slice is 100 and each register bit is comprised of 50 junctions. Therefore, for a 128-bit-deep, dual-bank register, the ratio of register junctions to processor junctions is 128:1. The fact that 99% of SFQ chip area is taken up with registers significantly simplifies vertical interconnect requirements. A large number of connections will be required between processor bit slices, and would, in turn, inject significant complexity into the 3DI

stacking scheme if the stacking includes processing units. If it is assumed, however, that fabrication technology commensurate with a circuit of this complexity exists (i.e. 8 – 10 levels of superconducting interconnects), then there is no need to stack processing elements and only the stacking of memory need be considered. An accounting of signal lines passing to the registers, stacked level by level is as follows: 3 sets of 7 bit wide address lines (2 read, 1 write), 64 instruction lines, three data bus lines (again, 2 read, 1 write), a clock line and two handshake lines, for a total of 91 which is rounded to 100 for margin. The remaining chip area would be filled with current-carrying TSVs, and power dissipation from current delivery will drop as the square of the number of available TSVs.

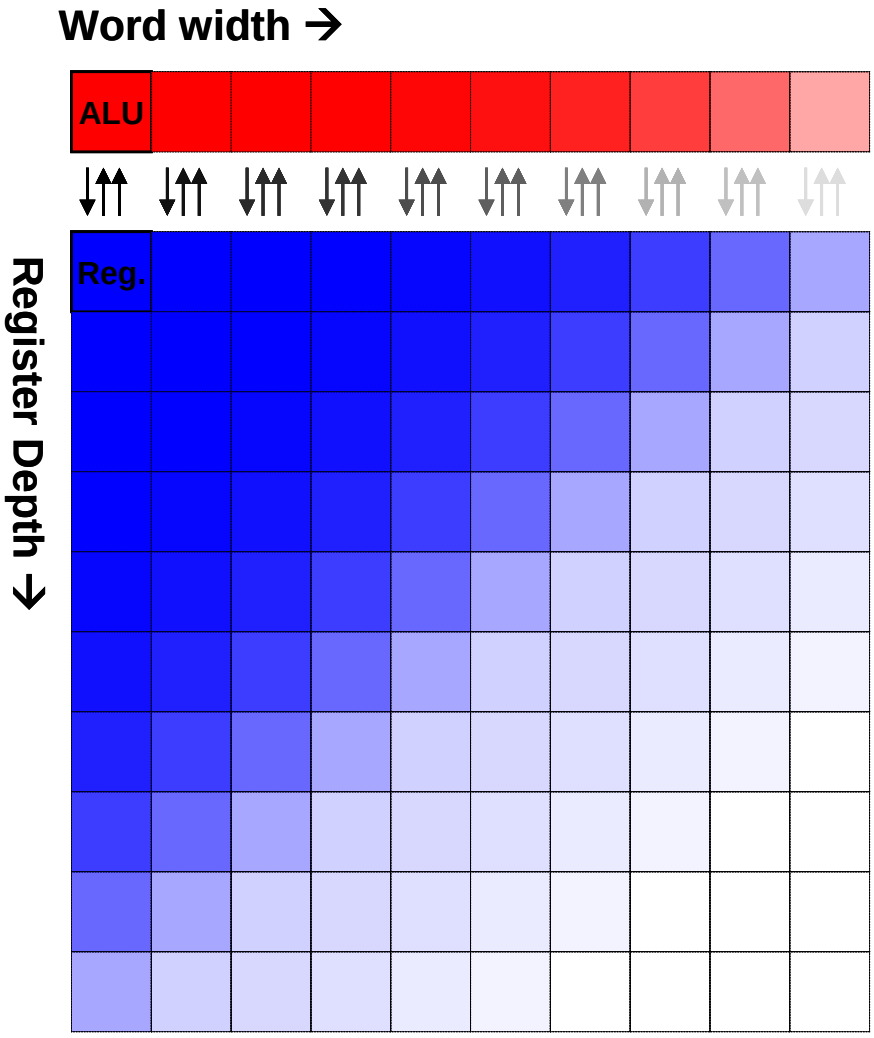

Figure 9: Notional Diagram of Processor

The first step in developing normal metal TSVs for 3DI interconnects is to determine whether an SFQ pulse can be effectively transmitted through such a structure. In all calculations which follow, it is assumed that the underlying SFQ technology is based upon junctions with $100\mu A/\mu m^2$ critical current density, a packing density of 10^5 junctions / cm^2, that the wafers through which the TSVs pass have been thinned to 100 μm and that via resistivity is 0.025 Ω·μm, that of CVD tungsten at 4K. Figure 10 shows

the inductance of an isolated TSV as a function of its diameter. Also graphed is the quantizing inductance as a function of its critical current.

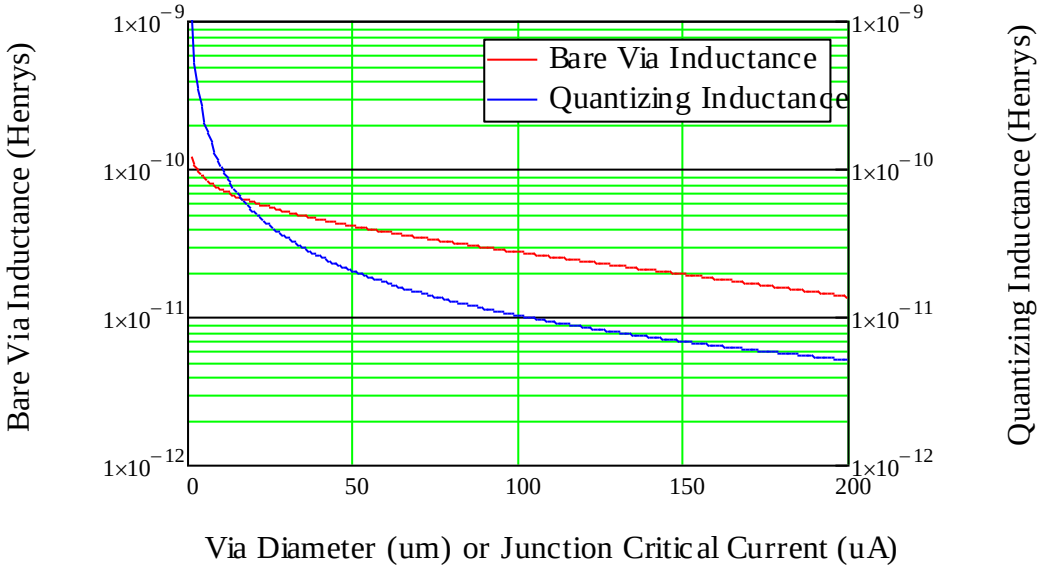

Figure 10: Bare Via and Quantizing Inductance vs. Via Diameter and Ic, Respectively

It is clear that the inductance of the via can not be made small compared to the quantizing inductance for reasonable values of junction critical current and via diameter. On the other hand, the TSV can be a quantizing inductance. Consider, for example, that a junction with a critical current of 50 μA has a quantizing inductance of 20 pH and that a 150 μm diameter TSV also has a 20 pH inductance. Unfortunately, using TSVs as quantizing inductances is impractical since, in order to be treated as truly isolated, the pitch would need to approach one mm, which would not be sufficiently dense. Brought closer together, inductance values would shift and cross-talk would become an issue, as well.

It is clear from the isolated example that via diameters will have to be of the order 100 μm in order for inductance to be held to reasonable values. Vias this wide cannot be effectively utilized since at this low aspect ratio, the fill material would have to be deposited to the full Si wafer thickness of 100 μm, leading to issues with film stress and extremely long polish times for planarizing the via fill. To circumvent this, the via can be comprised of a number of narrow slots, of critical dimension 3-5 μm which can then be filled conformally with only 2-3 μm of fill material. There are many patterns which can be used, and the pattern shown in Figure 11 has been chosen for analysis.

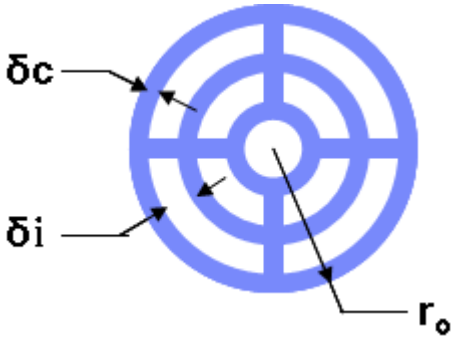

Figure 11: Cross-section of Through Silicon Via (TSV)

The TSV radius, r_o, is variable, while the conductor width, δ_c, and the insulator width, δ_c, are fixed at 5 μm each. Figure 12 shows the via resistance as a function of diameter. The discontinuities in the curve are the result of the quantized conductor and insulator widths. For diameters greater than 50 μm, the resistance is of order mΩ and will not impact SFQ pulse propagation.

In order to form viable TSVs, they will need to be shielded and reduced in inductance. One means of accomplishing this is to form a "caged" signal line whereby the signal propagates along a TSV which is surrounded by four grounded TSVs, as indicated in Figure 13. Grounded TSVs can be shared in order to maximize packing density.

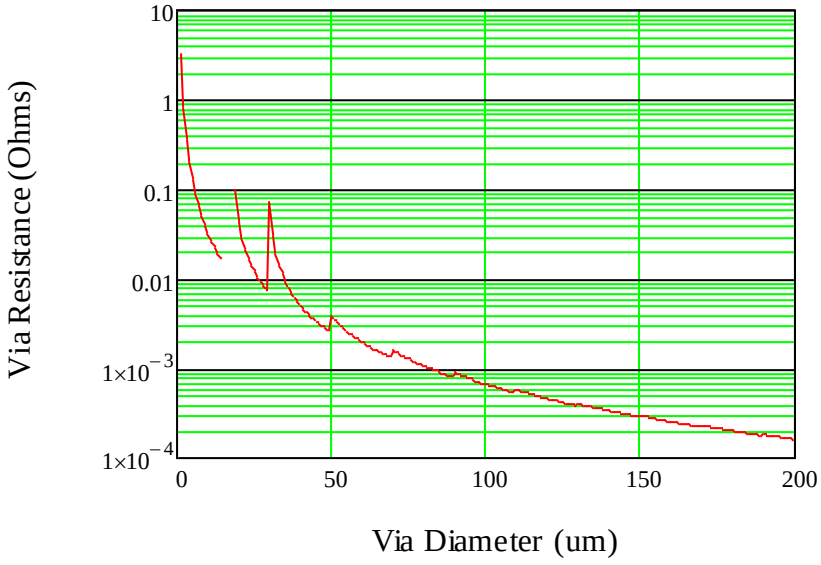

Figure 12: Via Resistance vs. Via Diameter

The TSV inductance in the presence of nearby vias is calculated using the formulae derived in[10], which properly takes into account the mutual inductance of closely spaced vias. Capacitance is calculated from the standard formula for parallel cylindrical

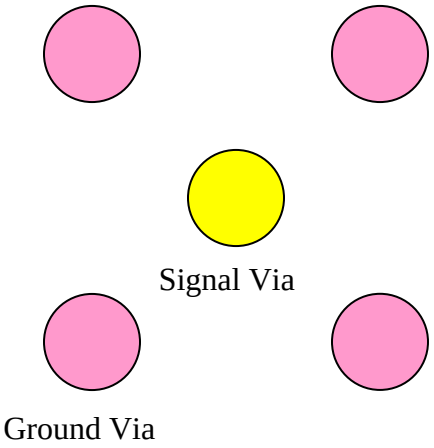

Figure 13: "Caged" Via Structure

conductors. In order to avoid reflected signals, the coaxial TSV must be impedance matched to the transmitting / receiving junctions. Figure 14 shows the impedance of a caged via for a range of diameters as a function of signal via to conductor via centerline spacing and the impedance of a Josephson junction as a function of its critical current. The graph indicates that for sufficiently low junction critical current and sufficiently large spacing, matched vias and junctions are possible. For example, a junction with a critical current of 170 μA is impedance matched to a 100 μm diameter via with 125 μm centerline spacing from the ground via. Decreasing junction critical current to 90 μA allows a match to 100 μm diameter vias with 200 μm centerline spacing.

Figure 15 plots caged via inductance as a function of signal-ground spacing and also quantizing inductance as a function of junction critical current. For parameter values which gave a matched case: I_c = 170 μA, via diameter = 100 μm and signal ground spacing = 125μm, the via inductance is 1/3 the quantizing inductance, which is good safety margin. For the case of a 90 μA junction and 100 μm diameter vias with 200 μm centerline spacing, the via inductance increases to ~50% of the quantizing inductance. These parameter values for the via diameter and spacing are quite reasonable for interfacing with solder bumps. For the case of junctions with 170 μA critical current, 63 μm bumps with 125 μm minimum spacing (i.e. following the same pattern as the caged vias with shared grounds) would be required and comprise a manufacturable configuration. Junctions with 90 μA critical current allow use of 100 μm bumps.

Current supply through TSVs is now considered. Figure 16 gives the number of vias available for current delivery as a function of the via diameter and number of stacked

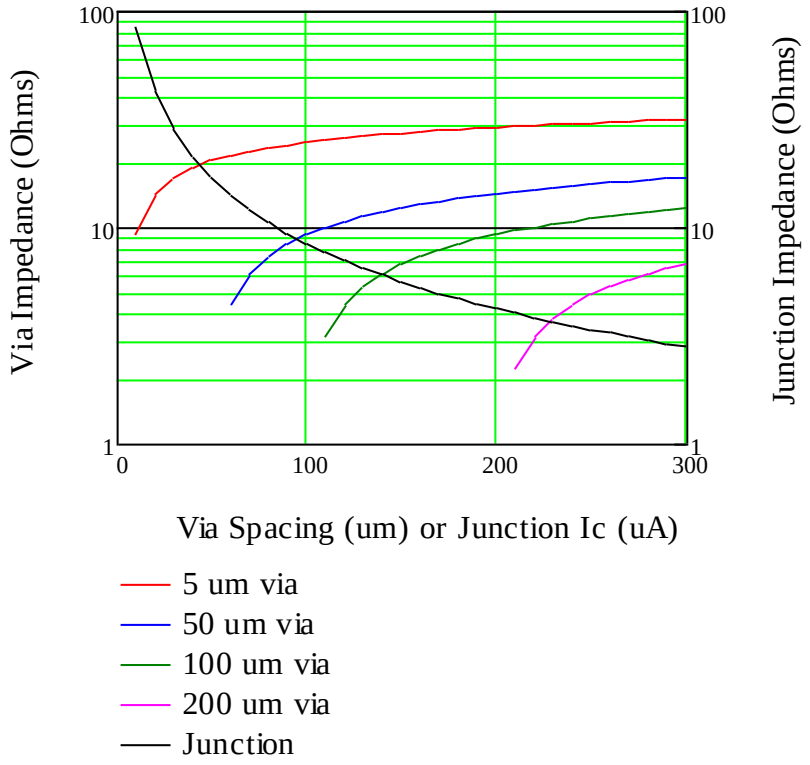

Figure 14: Via and Junction Impedance vs. Signal-Ground and Ic, Respectively

memory layers under the assumption that via pitch is twice the diameter and all chip area not taken up by signal TSVs is available for current delivery. For this plot, signal vias are 100 μm in diameter, spaced to match 170 μA junctions and are arranged in a contiguous block with shared grounds.

A final figure of merit which must be calculated is the power dissipated in the TSVs. The calculation assumes caged signal lines and subtracts from the total via count the signal and signal ground count (which is twice the signal via count). The remainder of the vias are assumed to carry bias current for the registers, with half used for supply and half for return. The number of stacked chips has a significant impact on power dissipation. For a fixed register file size, the number vias available in the chip's area is inversely proportional to the number of stacked layers, while the number of vias traversed in the direction normal to the chip is also proportional to the number of layers. Since power is dissipated as the square of the current, the dissipated power increases as the square of the number of stacked chips. This is evident in Figure 17, which gives a factor of 16 difference between the ratio of resistive power dissipated in vias to the circuit switching power dissipation as a function of the number of 64-bit registers sharing a serial biasing "ground island". Figure 17, which is plotted for an average critical current of 100 μA, a clock frequency of 10 GHz, and a current supply TSV diameter of 100 μm, also shows the value of serial biasing for both the reduction of total chip power (only the 64-bit wide register being accessed must switch) and for minimizing the fraction of chip power resistively dissipated in the TSVs. And since dynamic power dissipation is proportional to clock frequency, while bias current dissipation is constant, the fractional contribution of bias line power will vary inversely with the clock frequency.

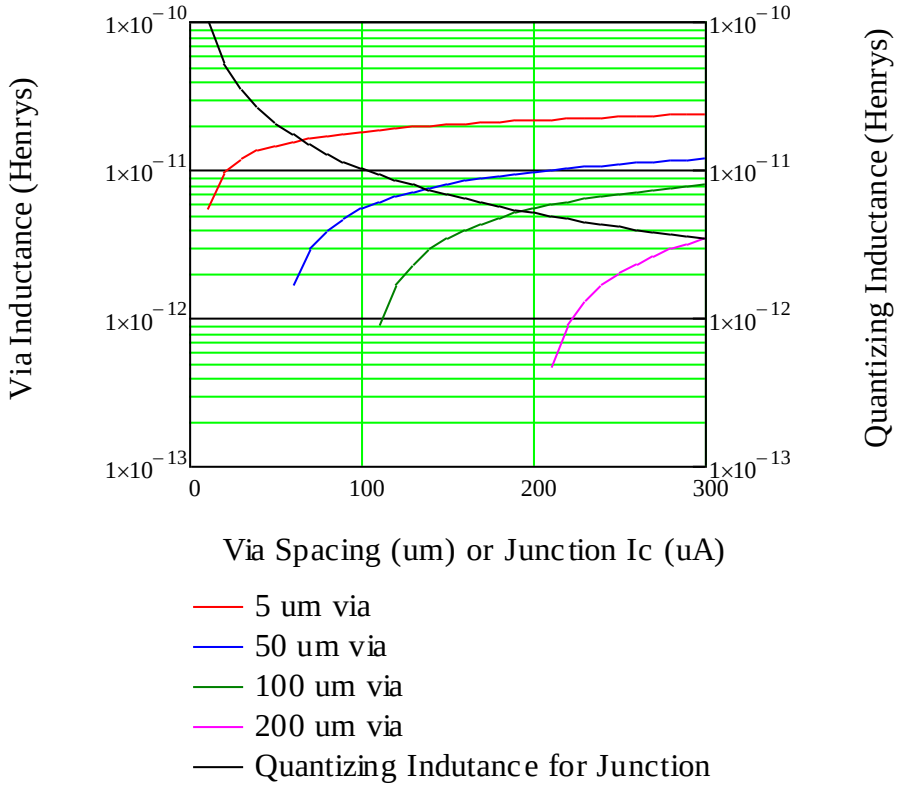

Figure 15: Via and Quantizing Inductance vs. Signal-Ground and Ic, Respectively

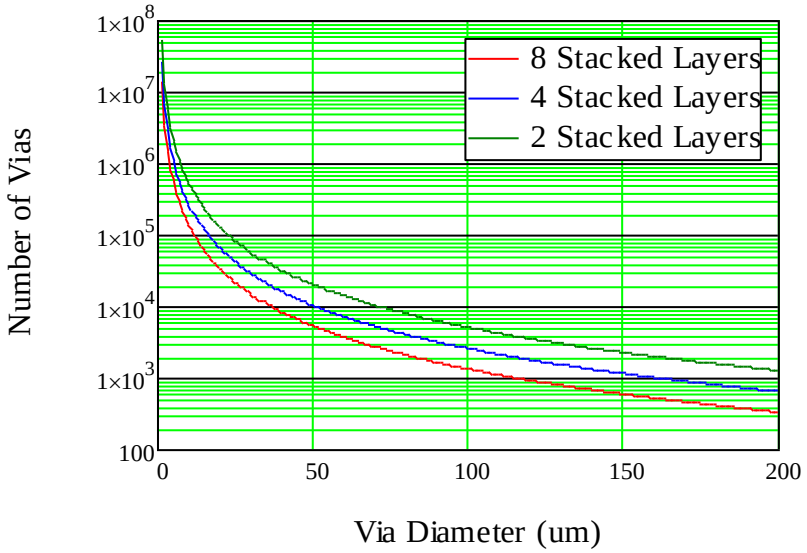

Figure 16: Number of Vias Available for Current Delivery

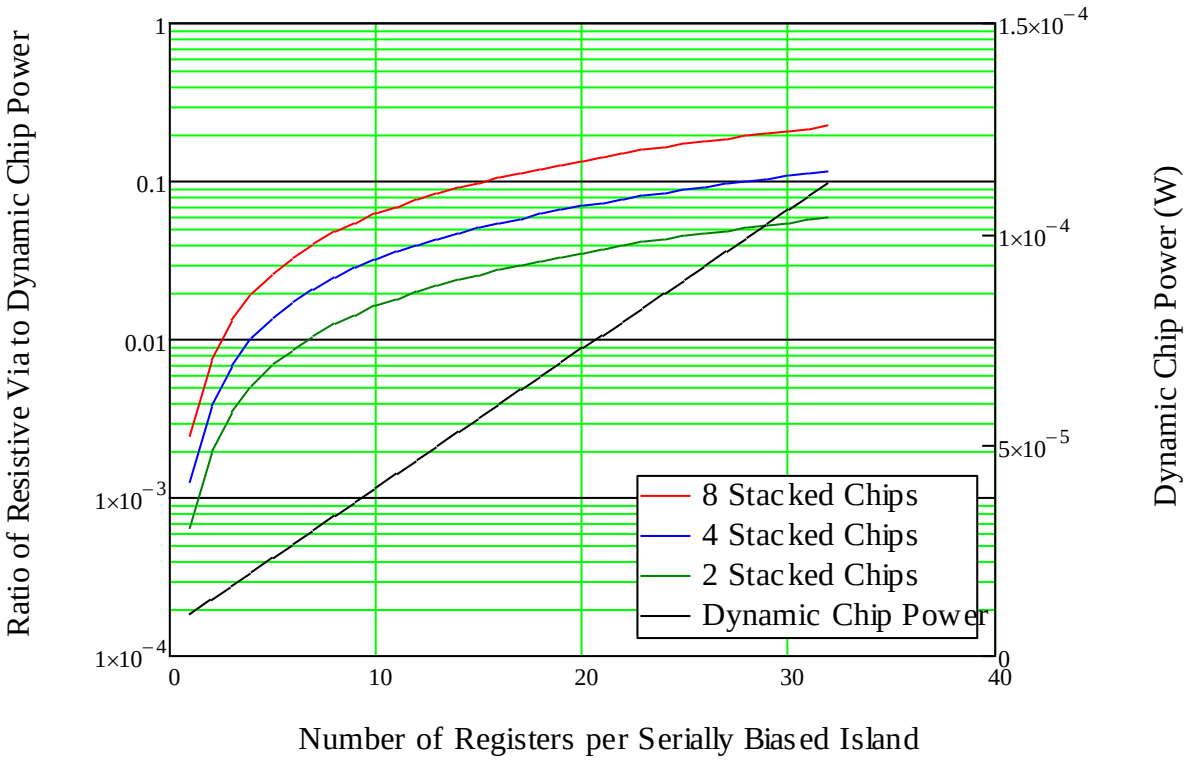

Figure 17: Ratio of Via to Active Circuit Power Dissipation

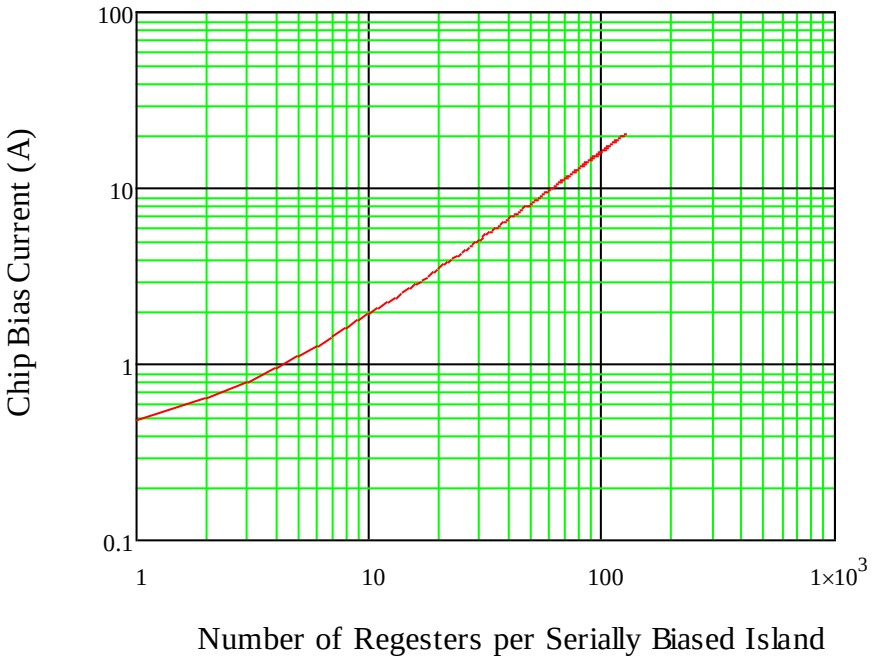

Figure 18: Chip Bias Current with Serial Biasing of Registers

Figure 18 plots the total chip current as a function of the number of registers on a serial biasing "island". The need to minimize the number of registers per island in order to keep total supplied current to a minimum is clear and is independent of the constraint imposed by ground shifts across the stacked layers imposed by IR drop across the TSVs.

In conclusion, we have shown that a Windsor Blue architecture which employs a cryogenic CMOS main memory, runs at a clock speed of 10 GHz, utilizes serial biasing of registers and employs stacked memory can meet density and power efficiency metrics which enable it to compete with the CMOS version.

References

[1]S. Polonsky, *IEEE Trans. App. Superconductivity* v9 n2 p3535 (1999)
[2]A. Rylyakov and S. Polonsky, *IEEE Trans. App. Superconductivity* v8 n1 p14 (1998)
[3]P. Patra *et al*, *Extended abstracts of ISEC '97* p42 (1997)
[4]"Design of all-dc-powered high-speed single flux quantum random access memory based on a pipeline structure for memory cell arrays", Nagasawa *et al,* Supercond. Sci. Tech. v19pS325 (2006)
[5]"Yield Evaluation of 10-kA/cm2 Nb Multi-Layer Fabrication Process Using Conventional Superconducting RAMs", Nagasawa *et al* , IEEE TRANSACTIONS ON APPLIED SUPERCONDUCTIVITY, v17n2p177 (2007)
[6]"Low Power Digital Design", Mark Horowitz, 1994 IEEE Symposium on Low Power Electronics, October 1994.
[7]"Data-Driven Self-Timed RSFQ Digital Integrated Circuit and System", Deng *et al*, IEEE TRANSACTIONS ON APPLIED SUPERCONDUCTIVITY, v7n2p3634 (1997)
[8]"Current Recycling and SFQ Signal Transfer in Large Scale RSFQ Circuits", J.H. Kang and S.B Kaplan, IEEE TRANSACTIONS ON APPLIED SUPERCONDUCTIVITY, v13n2p547 (2003)
[9]"SFQ Pulse Transfer Circuits Using Inductive Coupling for Current Recycling", Igarashi *et al*, IEEE TRANSACTIONS ON APPLIED SUPERCONDUCTIVITY, v19n3p649 (2009)
[10]"Identifying and Quantifying Printed Circuit Board Performance", Hubing, *et al*, IEEE International Symp on EMC p205 (1994)